CERTIFIED COPY OF
PRIORITY DOCUMENT

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２０００年　５月１７日

出願番号
Application Number: 特願２０００－１４５１６８

出願人
Applicant(s): 松下電器産業株式会社

CERTIFIED COPY OF
PRIORITY DOCUMENT

２００１年　５月１８日

特許庁長官
Commissioner,
Japan Patent Office

及川耕造

【書類名】　特許願

【整理番号】　2033820044

【提出日】　平成12年 5月17日

【あて先】　特許庁長官殿

【国際特許分類】　G06F 15/403

G06F 15/40

G06F 15/20

【発明者】

【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内

【氏名】　内藤　栄一

【発明者】

【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内

【氏名】　荒木　昭一

【発明者】

【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内

【氏名】　九津見　洋

【発明者】

【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内

【氏名】　小澤　順

【発明者】

【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内

【氏名】　丸野　進

【特許出願人】

【識別番号】　000005821

【氏名又は名称】　松下電器産業株式会社

【代理人】

【識別番号】　100097445

【弁理士】

【氏名又は名称】　岩橋　文雄

【選任した代理人】

【識別番号】　100103355

【弁理士】

【氏名又は名称】　坂口　智康

【選任した代理人】

【識別番号】　100109667

【弁理士】

【氏名又は名称】　内藤　浩樹

【手数料の表示】

【予納台帳番号】　011305

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9809938

【書類名】　明細書

【発明の名称】　情報検索装置、自動質問回答装置、情報検索方法および自動質問回答方法

【特許請求の範囲】

【請求項１】　検索対象の複数の情報単位から検索条件に合致する情報単位を検索する情報検索装置であって、検索対象の情報単位を記憶する情報単位記憶手段と、前記情報単位の特徴量を抽出する特徴量抽出手段と、前記特徴量に基づき前記情報単位のクラスタ分類を行うクラスタ分類手段と、前記情報単位から検索条件に合致する情報単位を検索する検索手段と、検索時に検索結果の情報単位が属する前記クラスタを出力する出力手段を有することを特徴とする情報検索装置。

【請求項２】　前記情報単位は、文章の項目を少なくとも1つ含むことを特徴とする請求項１に記載の情報検索装置。

【請求項３】　前記特徴量抽出手段は、前記情報単位の文章の項目から単語を切り出し、単語とその重みの組とからなる特徴量を抽出することを特徴とする請求項２に記載の情報検索装置。

【請求項４】　前記特徴量に基づき前記クラスタの内容を表すラベルを作成するクラスタラベル作成手段をさらに有し、前記出力手段は検索結果の情報単位が属する前記クラスタの前記ラベルを出力することを特徴とする請求項1に記載の情報検索装置。

【請求項５】　前記クラスタラベル作成手段は、前記特徴量に基づき前記クラスタの内容を表す単語からなるラベルを作成することを特徴とする請求項４に記載の情報検索装置。

【請求項６】　前記クラスタラベル作成手段は、前記特徴量に基づき前記クラスタの内容を表す文からなるラベルを作成することを特徴とする請求項４に記載の情報検索装置。

【請求項７】　前記特徴量に基づき前記情報単位の内容を表すラベルを作成する情報単位ラベル作成手段をさらに有し、前記出力手段は検索結果の情報単位の前記ラベルと該情報単位が属するクラスタの前記ラベルとを出力することを特徴

とする請求項４に記載の情報検索装置。

【請求項８】　情報単位ラベル作成手段は、前記特徴量に基づき前記クラスタの内容を表す文からなるラベルを作成することを特徴とする請求項７に記載の情報検索装置。

【請求項９】　検索対象の複数の情報単位から検索条件に合致する情報単位を検索する情報検索方法であって、前記情報単位の特徴量を抽出し、前記特徴量に基づき前記情報単位のクラスタ分類を行い、前記クラスタの内容を表すラベルを作成し、前記特徴量に基づき前記情報単位の内容を表す文からなるラベルを作成し、前記情報単位から検索条件に合致する情報単位を検索し、検索結果の情報単位の前記ラベルと該情報単位が属するクラスタの前記ラベルとを出力することを特徴とする情報検索方法。

【請求項１０】　複数の文書から利用者の質問に合致する文書を検索し利用者に回答する自動質問回答装置であって、回答対象の文書を記憶する文書記憶手段と、前記文書の特徴量を抽出する特徴量抽出手段と、前記特徴量に基づき前記文書のクラスタ分類を行うクラスタ分類手段と、前記特徴量に基づき前記クラスタの内容を表すラベルを作成するクラスタラベル作成手段と、前記特徴量に基づき前記文書の内容を表すラベルを作成する文書ラベル作成手段と、利用者からの質問と前記文書との類似度を算出する類似度算出手段と、前記類似度が所定の値以上の文書を回答文書として選択する回答選択手段と、前記回答文書の前記ラベルと該回答文書が属するクラスタの前記ラベルとを出力する出力手段とを有することを特徴とする自動質問回答装置。

【請求項１１】　複数の文書から利用者の質問に合致する文書を検索し利用者に回答する自動質問回答方法であって、前記文書の特徴量を抽出し、前記特徴量に基づき前記文書のクラスタ分類を行い、前記特徴量に基づき前記クラスタの内容を表すクラスタラベルを作成し、前記特徴量に基づき前記文書の内容を表すラベルを作成し、利用者からの質問と前記文書との類似度を算出し、前記類似度が所定の値以上の文書を回答文書として選択し、前記回答文書の前記ラベルと該回答文書が属するクラスタの前記ラベルとを出力することを特徴とする自動質問回答方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、大量の情報の中から利用者が求める情報を容易に見つけ出すことを可能とするための情報検索装置および、利用者からの質問に対して適切な回答を検索して回答する自動質問回答装置に関する。

【０００２】

【従来の技術】

近年、インターネットの普及に伴い、ＷＷＷ（World Wide Web）上にＨＴＭＬ（Hyper Text Markup Language）で記述された様々なホームページが掲載されるようになるなど、一般利用者が大量の情報にアクセスすることが可能となっている。また、ＦＡＱ(Frequently Asked Questions)集と称した、頻繁に問い合わせられる質問とその回答とを対にしたリストが公開されていて、利用者は質問に対する回答を得ることが可能である。これらの情報は、利用者にとって、求める情報の所在がわかれば、即座に閲覧できるので便利であるが、逆に大量の情報の中から自分の求める情報を見つけ出すことが大変な作業となっている。

【０００３】

このため、文書からキーワードを切り出してその文書の特徴量とし、特徴量間の内積を算出して文書間の類似度を求め、質問文に対する類似文書を検索するという検索技術が知られている。

【０００４】

【発明が解決しようとする課題】

しかしながら、インターネット上の情報や、あるいは、事例ベースで蓄積されたＦＡＱ集は、多くの人が独立して情報を提供しているので、情報の重複は避けられず、同じような内容を持つ文書が大量に存在する。したがって、従来の技術では、質問文に類似した文書として、同じような内容の文書が大量に検索されてしまうことが多いので、利用者は結局大量の検索結果の中から欲しい情報を見つけ出す作業が必要であった。もしくは、検索結果を一定の数に制限すると、同じような内容の検索結果で占められてしまい、自分の欲しい情報がなかったりする

という課題があった。

【０００５】

本発明は、かかる点に鑑み、同じような内容の検索結果をまとめて出力することにより、利用者に検索結果の内容を把握することを容易にし、情報検索の負担を軽減させる情報検索装置、自動質問回答装置、情報検索方法および自動質問回答方法を提供することを目的とする。

【０００６】

【課題を解決するための手段】

請求項１の本発明は、検索対象の複数の情報単位から検索条件に合致する情報単位を検索する情報検索装置であって、検索対象の情報単位を記憶する情報単位記憶手段と、前記情報単位の特徴量を抽出する特徴量抽出手段と、前記特徴量に基づき前記情報単位のクラスタ分類を行うクラスタ分類手段と、前記情報単位から検索条件に合致する情報単位を検索する検索手段と、検索時に検索結果の情報単位が属する前記クラスタを出力する出力手段を有することを特徴とする情報検索装置である。

【０００７】

請求項４の本発明は、前記特徴量に基づき前記クラスタの内容を表すラベルを作成するクラスタラベル作成手段をさらに有し、前記出力手段は検索結果の情報単位が属する前記クラスタの前記ラベルを出力することを特徴とする請求項1に記載の情報検索装置である。

【０００８】

請求項７の本発明は、前記特徴量に基づき前記情報単位の内容を表すラベルを作成する情報単位ラベル作成手段をさらに有し、前記出力手段は検索結果の情報単位の前記ラベルと該情報単位が属するクラスタの前記ラベルとを出力することを特徴とする請求項４に記載の情報検索装置である。

【０００９】

請求項９の本発明は、複数の文書から利用者の質問に合致する文書を検索し利用者に回答する自動質問回答装置であって、回答対象の文書を記憶する文書記憶手段と、前記文書の特徴量を抽出する特徴量抽出手段と、前記特徴量に基づき前

記文書のクラスタ分類を行うクラスタ分類手段と、前記特徴量に基づき前記クラスタの内容を表すラベルを作成するクラスタラベル作成手段と、前記特徴量に基づき前記文書の内容を表すラベルを作成する文書ラベル作成手段と、利用者からの質問と前記文書との類似度を算出する類似度算出手段と、前記類似度が所定の値以上の文書を回答文書として選択する回答選択手段と、前記回答文書の前記ラベルと該回答文書が属するクラスタの前記ラベルとを出力する出力手段とを有することを特徴とする自動質問回答装置である。

【００１０】

【発明の実施の形態】

以下、本発明の実施の形態について図面を参照しながら説明する。

【００１１】

（実施の形態１）

第１の実施の形態の機能構成図を図１に示す。図１において、１０１は複数の文書を記憶する文書記憶部、１０２は文書記憶部１０１に記憶されている文書から特徴ベクトルを抽出する特徴ベクトル抽出部、１０３は特徴ベクトル抽出部１０２が求めた特徴ベクトルに基づき文書記憶部１０１に記憶されている文書のクラスタ分類を行うクラスタ分類部、１０４はクラスタ分類部１０３がクラスタ分類した文書のクラスタを記憶するクラスタ記憶部、１０５は文書記憶部１０１とクラスタ記憶部１０４とから文書を検索するデータベース検索部、１０６は利用者との入出力を管理する利用者入出力部、１０７は利用者からの入力を受け付ける入力部、１０８は利用者に情報を表示する表示部である。

【００１２】

本実施の形態のハードウェア構成を図２に示す。図２は基本的に汎用の計算機システムの構成と同じである。また、図１に示した構成部分と同一の構成部分を含んでいるため、同一構成部分には同一番号を付して説明を省略する。図２において、１１２はプログラムやデータを記憶する揮発性メモリからなる主記憶装置、１１３はプログラムやデータを記憶する不揮発性メモリからなる補助記憶装置、１１１は主記憶装置１１２に記憶されているプログラムを実行するＣＰＵである。補助記憶装置１１３に記憶されているプログラムやデータは、主記憶装置１

１２にロードされた後ＣＰＵ１１１により実行される。

【００１３】

文書記憶部１０１には、検索の対象となる所与のn個の文書が記憶されている(n≧2)。記憶されている文書の例を図３に示す。文書は、ユニークな文書ID、文章形式の本文とからなる。i番目の文書をD_iとする(1≦i≦n)。

【００１４】

以上のように構成された第１の実施の形態の動作を、文書登録時と文書検索時に分けて以下に説明する。文書登録時とは、初めて文書が文書記憶部１０１に登録される場合、あるいは、それ以降に文書の追加／変更／削除があった場合に行われる動作である。文書検索時とは、登録されている文書を検索して閲覧する場合に行われる動作である。

【００１５】

「文書登録時」

まず、特徴ベクトル抽出部１０２は、文書記憶部１０１に記憶されている全ての文書D_iを取り出し、各文書の特徴ベクトルV_iを抽出する。特徴ベクトルは、文書の特徴を表す単語t_jとその重みw_{ij}の組を要素とするベクトルであり、その要素の数は文書によって異なる。ここで、jは単語を識別するユニークな番号である。抽出された特徴ベクトルV_iの例を図４に示す。特徴ベクトルの抽出手順を図５に示すフローチャートを用いて説明する。

【００１６】

[ステップ101]　カウンタの初期化

文書のカウンタiにi=1を設定する。

【００１７】

[ステップ102]　単語の抽出

文書記憶部１０１から文書D_iを取り出し、形態素解析、構文解析、不要語除去など、一般に知られている方法により、本文から出現する単語t_jを抽出し、t_jの文書内での出現回数f_{ij}をカウントする。

【００１８】

[ステップ103]　終了判定

全文書につき[ステップ102]の処理が終了した場合、すなわちi=nの場合[ステップ105]に進む。そうでない場合[ステップ104]に進む。

【００１９】

[ステップ104]　カウンタの増加

カウンタiを1増加し[ステップ102]に進む。

【００２０】

[ステップ105]　idf値算出

単語t_jの全文書に対する重要度として、単語t_jが出現する文書数の少なさを表す度合idf(inverse document frequency)値を（数１）により算出する。

【００２１】

【数１】

$$idf_j = \log \frac{n}{\text{語}t_j\text{が出現する文書数}} + 1$$

【００２２】

[ステップ106]　カウンタの初期化

文書のカウンタiにi=1を設定する。

【００２３】

[ステップ107]　tf･idf値算出

単語t_jが文書D_iを特徴付ける重みw_{ij}として、単語t_jの文書D_i内での出現割合tf(term frequency)値とidf値とをかけ合わせたtf･idf値を（数２）により算出する。

【００２４】

【数２】

$$w_{ij} = \frac{f_{ij}}{\sum_{j:t_j \in D_i} f_{ij}} \cdot idf_j$$

【００２５】

[ステップ108]　終了判定

全文書につき[ステップ107]の処理が終了した場合、すなわちi=nの場合終了する。そうでない場合[ステップ109]に進む。

【００２６】

[ステップ109]　カウンタの増加

カウンタiを1増加し[ステップ102]に進む。

【００２７】

次に、クラスタ分類部１０３は、特徴ベクトル抽出部１０２が抽出した特徴ベクトルを用いて、全ての文書をm個のクラスタに分類する(1<m<n)。ここで、k番目のクラスタをC_kとする(1≦k≦m)。クラスタ分類の手順として、樹形図的に逐次クラスタに分類していく階層的クラスタリングを用い、図６を用いて説明する。

【００２８】

[ステップ111]　クラスタ間初期距離計算

クラスタの初期値として、１つの文書D_iだけを要素として持つクラスタ$C_i=[D_i]$を設定する。すなわち初期クラスタはn個ある。各クラスタC_k,C_l(1≦k,l≦n)間の距離d_{kl}として、各文書の特徴ベクトル間の距離を類似比を用いて（数３）により算出する。（数３）において、∧はmin演算、∨はmax演算を表す。

【００２９】

【数３】

$$d_{kl} = -\log \frac{\sum_{j:t_j \in D_k \cup D_l} w_{kj} \wedge w_{lj}}{\sum_{j:t_j \in D_k \cup D_l} w_{kj} \vee w_{lj}}$$

【００３０】

[ステップ112]　カウンタの初期化

クラスタリング回数のカウンタiにi=1を設定する。

【００３１】

[ステップ113]　距離最小クラスタ探索

全てのクラスタの組み合わせの中で、クラスタ間距離D_{kl}が最も小さいクラスタC_k,C_l(k<l)の組を探索する。

【００３２】

[ステップ114]　クラスタ統合

クラスタC_k,C_lを統合してクラスタC_gとする。すなわち、$C_g=C_k \cup C_l$、$C_l=\phi$とする(ϕは空集合を表す)。クラスタの統合に伴い、クラスタC_gと他のクラスタC_h(1≦h≦n)とのクラスタ間距離をウォード法を用いて（数４）により算出する。（数４）においてn_kはクラスタC_kの要素の数である。

【００３３】

【数４】

$$d_{gh}=\frac{(n_k+n_h)\cdot d_{kh}+(n_l+n_h)\cdot d_{lh}-n_h\cdot d_{kl}}{n_g+n_h}$$

【００３４】

[ステップ115]　終了判定

クラスタリング回数がn-1の場合、すなわち全ての初期クラスタが1つのクラスタに統合された場合[ステップ117]に進む。そうでない場合[ステップ116]に進む。

【００３５】

[ステップ116]　カウンタの増加

カウンタiを1増加し[ステップ112]に進む。

【００３６】

[ステップ117]　クラスタ数決定

[ステップ111]から[ステップ115]までのクラスタ分類過程においては、クラスタリング回数ごとにクラスタの数は1つずつ減少する。このステップでは、クラスタ分類過程を振り返り、適切なクラスタリング回数を決定する。ここでは、要素を2つ以上持つクラスタの数が最大になるクラスタリング回数を適切なクラス

タリング回数であるとする。

【００３７】

[ステップ118]　クラスタ要素書き出し

[ステップ117]で決定したクラスタリング回数までクラスタ分類を行った時点での各クラスタに含まれる要素をクラスタ記憶部１０４に書き出す。クラスタ記憶部１０４に書き出されたクラスタの例を図７に示す。各クラスタは、クラスタIDとそのクラスタに含まれる文書の文書IDとからなる。例えば、クラスタ1には、1，190，432，644番の4つの文書が含まれている。これは、この4つの文書の特徴ベクトル同士が、他の文書にくらべて類似していることを表している。

【００３８】

以上の動作により、文書登録時に、文書のクラスタを作成して記憶しておく。

【００３９】

「文書検索時」

まず、利用者入出力部１０６は、入力部１０７を通じて文書の検索条件を受け付ける。検索条件としては、文書のキーワード、文書IDなど文書検索の条件となるものなら何でもよい。

【００４０】

データベース検索部１０５は、文書記憶部１０１から検索条件を満たす文書を検索し、次に、クラスタ記憶部１０４から検索された文書が含まれるクラスタを検索し、最後に、再び文書記憶部１０１から検索されたクラスタに含まれる文書を検索し、その結果を利用者入出力部１０６に送る。

【００４１】

利用者入出力部１０６は、表示部１０８を通じて検索結果を利用者に表示する。検索結果の例を図８に示す。図８では、クラスタIDと、そのクラスタに含まれる文書の文書IDと本文とを、クラスタごとに表形式で表示し、マウスで「前のクラスタ」ボタンや「次のクラスタ」ボタンを押して別のクラスタを表示することにより、全ての検索結果を表示することができる。

【００４２】

以上のように、第１の実施の形態によれば、文書の特徴ベクトルを抽出し、特

徴ベクトルに基づいて文書をクラスタ分類して記憶し、文書の検索結果をクラスタごとにまとめて表示することにより、検索結果を類似した文書の固まりとして把握することが容易となるので、利用者の文書検索の負担を軽減することが可能となる。

【００４３】

なお、本実施の形態では、文書は所与のものがあらかじめ記憶されていたが、光ディスクなどの記憶媒体やインターネットなどのネットワーク媒体などにより、後から新たに導入、もしくは、改訂されても良い。また、対象情報特徴ベクトルの算出は、tf･idf値を用いていたが、単純に単語の出現回数とするなど、他の方法でも良い。また、クラスタ分類の方法として、階層的クラスタリングを用いたが、非階層的クラスタリングでも良い。また、初期クラスタ間距離として（数３）の類似比を用いたが、ユークリッド平方距離など他の距離を用いても良い。また、クラスタ統合時のクラスタ間距離の算出手法として（数４）のウォード法を用いたが、最長距離法など他の手法を用いても良い。また、クラスタ数の決定手法として、要素を2つ以上持つクラスタの数が最大になるクラスタリング回数としたが、クラスタ数を文書数の一定の割合とするなど他の決定手法でも良い。また、文書の検索は、キーワードや文書IDによるもの以外に、全文検索であってもあいまい検索であっても良い。また、検索結果として文書IDをも表示したが、表示しなくても良い。

【００４４】

（実施の形態２）

次に本発明の第２の実施の形態について説明する。第１の実施の形態では、検索結果の文書をクラスタにまとめて表示する例で説明したが、第２の実施の形態は、クラスタに単語からなる単語ラベルを付けて、クラスタの内容を把握し易くできるように考慮したものである。

【００４５】

第２の実施の形態の機能構成図を図９に示す。図９において、図１に示した第１の実施の形態と同じ構成部分には同一番号を付して説明は省略する。第２の実施の形態と第１の実施の形態との相違点は、クラスタ分類部１０３が作成したク

ラスタについて単語からなる単語ラベルを作成するクラスタ単語ラベル作成部２０１と、クラスタ単語ラベル作成部２０１が作成した単語ラベルを記憶するクラスタラベル記憶部２０２とを追加したことである。

【００４６】

文書記憶部１０１に記憶されている文書の例として、第１の実施の形態と同じ図３に示したものを用いる。

【００４７】

以上のように構成された第２の実施の形態の動作を、第１の実施の形態と同様に、文書登録時と文書検索時に分けて以下に説明する。

【００４８】

「文書登録時」

クラスタ分類部１０３がクラスタ分類を行うまでの処理は、第１の実施の形態の動作と同様であるので説明は省略する。それ以降のクラスタ単語ラベル作成部２０１の動作を図１０のフローチャートを用いて説明する。

【００４９】

[ステップ201]　カウンタの初期化

クラスタのカウンタkにk=1を設定する。

【００５０】

[ステップ202]　単語出現回数計数

クラスタC_kの要素である全ての文書D_iに含まれる単語t_jごとに、単語t_jが含まれる文書の数をカウントする。すなわち、単語t_jのクラスタC_k内での共起回数を数える。

【００５１】

[ステップ203]　tf･idf値積算

クラスタC_kの要素である全ての文書D_iに含まれる単語t_jごとに、単語t_j のtf･idf値w_{ij}の合計を算出する。

【００５２】

[ステップ204]　単語のソート

クラスタC_kの要素である全ての文書D_iに含まれる全ての単語t_jを、[ステップ2

02]で求めた共起回数の多い順にソートする。回数が同じ場合は[ステップ202]で求めたtf･idf値の合計の大きい順にソートする。

【００５３】

[ステップ205]　クラスタ単語ラベルの書き出し

[ステップ204]でソートされた上位の３つの単語を選択し、クラスタの単語ラベルとしてクラスタ記憶部２０２に書き出す。クラスタラベル記憶部に書き出された単語ラベルの例を図１１に示す。例えば、クラスタ1には、「お菓子」「間食」「チーズ」という単語ラベルが付いていることを表す。

【００５４】

[ステップ206]　終了判定

全クラスタにつき[ステップ202]から[ステップ205]までの処理が終了した場合、すなわちk=mの場合終了する。そうでない場合[ステップ207]に進む。

【００５５】

[ステップ207]　カウンタの増加

カウンタkを1増加し[ステップ202]に進む。

【００５６】

以上の動作により、文書登録時に、文書のクラスタ、クラスタの単語ラベルとを作成して記憶しておく。

【００５７】

「文書検索時」

データベース検索部１０５がクラスタに含まれる文書を検索するまでの処理は、第１の実施の形態の動作と同様であるので説明は省略する。クラスタに含まれる文書を検索した後、データベース検索部１０５は、クラスタラベル記憶部２０２から、検索されたクラスタの単語ラベルを検索し、その結果を利用者入出力部１０６と表示部１０８を通じて利用者に表示する。検索結果の例を図１２に示す。図１２と第１の実施の形態での表示例の図８との違いは、クラスタにクラスタ単語ラベルが表示されていることである。

【００５８】

以上のように、第２の実施の形態によれば、文書の特徴ベクトルを抽出し、特

徴ベクトルに基づいて文書をクラスタ分類し、単語からなるクラスタの単語ラベルを作成して記憶し、文書の検索結果をクラスタごとにまとめて単語ラベルとともに表示することにより、検索結果を類似した文書の固まりとして把握し、なおかつ、固まりの内容を理解することが容易となるので、利用者の文書検索の負担を軽減することが可能となる。

【００５９】

なお、本実施の形態では、単語ラベルの作成方法として単語の共起回数でソートしたが、tf･idf値のみでソートするなど他の方法でも良い。また、単語ラベルの単語数を３つにしたが、３つ以外でも良い。また、検索結果としてクラスタID、文書IDをも表示したが、表示しなくても良い。

【００６０】

（実施の形態３）

次に本発明の第３の実施の形態について説明する。第２の実施の形態では、単語からなるクラスタの単語ラベルを表示する例で説明したが、第３の実施の形態は、クラスタに文からなる文ラベルを付けて、クラスタの内容を自然文で把握し易くできるように考慮したものである。

【００６１】

第３の実施の形態の機能構成図を図１３に示す。図１３において、図９に示した第２の実施の形態と同じ構成部分には同一番号を付して説明は省略する。第３の実施の形態と第２の実施の形態との相違点は、クラスタ単語ラベル作成部２０１を、クラスタ分類部１０３が作成したクラスタについて文からなる文ラベルを作成するクラスタ文ラベル作成部３０１に置き換えたことである。

【００６２】

文書記憶部１０１に記憶されている文書の例として、第１の実施の形態と同じ図３に示したものを用いる。

【００６３】

以上のように構成された第３の実施の形態の動作を、第１の実施の形態と同様に、文書登録時と文書検索時に分けて以下に説明する。

【００６４】

「文書登録時」

クラスタ分類部１０３がクラスタ分類を行うまでの処理は、第１の実施の形態の動作と同様であるので説明は省略する。それ以降のクラスタ文ラベル作成部３０１の動作を図１４のフローチャートを用いて説明する。

【００６５】

[ステップ301]　カウンタの初期化

クラスタのカウンタkにk=1を設定する。

【００６６】

[ステップ302]　単語出現回数計数

第２の実施の形態と同様に、クラスタC_kの要素である全ての文書D_iに含まれる単語t_jごとに、単語t_jが含まれるクラスタ内の文書の数をカウントする。すなわち、単語t_jのクラスタC_k内での共起回数を数える。

【００６７】

[ステップ303]　文ごとに積算

クラスタC_kの要素である全ての文書D_iを構成する文ごとに、[ステップ302]でカウントした単語の共起回数の合計を算出する。ここで、文とは、文書を「。」などの句点で区切った1つ1つをいう。

【００６８】

[ステップ304]　文のソート

クラスタC_kの要素である全ての文書D_iを構成する文を[ステップ303]で求めた共起回数の合計の大きい順にソートする。

【００６９】

[ステップ305]　クラスタ文ラベルの書き出し

[ステップ304]でソートされた最上位の文を選択し、クラスタの文ラベルとしてクラスタ記憶部２０２に書き出す。クラスタラベル記憶部２０２に書き出された文ラベルの例を図１５に示す。例えば、クラスタ1には、「水分の多い物（ゼリー、プリン、ヨーグルト）を…」という文ラベルが付いていることを表す。

【００７０】

[ステップ306]　終了判定

全クラスタにつき[ステップ302]から [ステップ305]までの処理が終了した場合、すなわちk=mの場合終了する。そうでない場合[ステップ307]に進む。

【００７１】

[ステップ307]　カウンタの増加

カウンタkを1増加し[ステップ202]に進む。

【００７２】

以上の動作により、文書登録時に、文書のクラスタ、クラスタの文ラベルとを作成して記憶しておく。

【００７３】

「文書検索時」

データベース検索部１０５が検索を行って検索結果を利用者に表示するまでの動作は、第２の実施の形態と同様である。検索結果の例を図１６に示す。図１６と第２の実施の形態での表示例の図１２との違いは、クラスタのラベルが単語ではなく文で表現されていることである。

【００７４】

以上のように、第３の実施の形態によれば、文書の特徴ベクトルを算出し、特徴ベクトルに基づいて文書をクラスタ分類し、文からなるクラスタの文ラベルを作成して記憶し、文書の検索結果をクラスタごとにまとめて文ラベルとともに表示することにより、検索結果を類似した文書の固まりとして把握し、なおかつ、固まりの内容を自然文として理解することが容易となるので、利用者の文書検索の負担を軽減することが可能となる。

【００７５】

なお、本実施の形態では、文ラベルの作成方法として単語の共起回数を合計したが、tf･idf値を合計するなど他の方法でも良い。また、単語ラベルの単語数を３つにしたが、３つ以外でも良い。また、検索結果としてクラスタID、文書IDをも表示したが、表示しなくても良い。

【００７６】

（実施の形態４）

次に本発明の第４の実施の形態について説明する。第３の実施の形態では、ク

ラスタの文ラベルを表示する例で説明したが、第４の実施の形態は、クラスタの文ラベルだけではなく、クラスタの要素である文書にも文による文書ラベルを付けて、文書の内容をも自然文で把握し易くできるように考慮したものである。

【００７７】

第４の実施の形態の機能構成図を図１７に示す。図１７において、図１３に示した第３の実施の形態と同じ構成部分には同一番号を付して説明は省略する。第４の実施の形態と第３の実施の形態との相違点は、クラスタ分類部１０３が作成したクラスタの要素である文書について文からなる文書ラベルを作成する文書ラベル作成部４０１と、文書ラベル作成部４０１が作成した文書ラベルを記憶する文書ラベル記憶部４０２を追加したことである。

【００７８】

文書記憶部１０１に記憶されている文書の例として、第１の実施の形態と同じ図３に示したものを用いる。

【００７９】

以上のように構成された第４の実施の形態の動作を、第１の実施の形態と同様に、文書登録時と文書検索時に分けて以下に説明する。

【００８０】

「文書登録時」

クラスタ文ラベル作成部３０１がクラスタの文ラベルを作成してクラスタラベル記憶部に書き出すまでは、第３の実施の形態の動作と同様であるので説明は省略する。それ以降の文書ラベル作成部４０１の動作を図１８のフローチャートを用いて説明する。

【００８１】

[ステップ401]　カウンタの初期化

文書のカウンタiにi=1を設定する。

【００８２】

[ステップ402]　tf･idf値積算

文書D_iを構成する各文ごとに、その文に含まれる全単語t_jのtf･idf値w_{ij}の合計を算出する。

【００８３】

[ステップ403]　終了判定

全文書につき[ステップ402]の処理が終了した場合、すなわちi=nの場合終了する。そうでない場合[ステップ404]に進む。

【００８４】

[ステップ404]　カウンタの増加

カウンタiを1増加し[ステップ402]に進む。

【００８５】

[ステップ405]　カウンタの初期化

クラスタのカウンタkにk=1を設定する。

【００８６】

[ステップ406]　文のソート

クラスタC_kの要素である全ての文書D_iを構成する文を、[ステップ402]で求めた合計の多い順にソートする。

【００８７】

[ステップ407]　文書ラベル選択

文書D_iの文書ラベルとして[ステップ406]でソートされた最上位の文を選択する。ただし、選択された文が、クラスタ文ラベル作成部が作成したクラスタの文ラベルと同一の場合は、文書D_iの文書ラベルとして[ステップ406]でソートされた上位から２番目の文を選択する。

【００８８】

[ステップ408]　文書ラベルの書き出し

[ステップ406]で選択された文書D_iの文ラベルを文書ラベル記憶部４０２に書き出す。文書ラベル記憶部４０２に書き出された文書ラベルの例を図１９に示す。例えば、クラスタ1に含まれる文書1には、「かみごたえがあり、後を引かないもので、…」という文書ラベルが付いていることを表す。

【００８９】

[ステップ409]　終了判定

全クラスタにつき[ステップ406]から [ステップ408]までの処理が終了した場

合、すなわちk=mの場合終了する。そうでない場合[ステップ410]に進む。

【００９０】

[ステップ410]　カウンタの増加

カウンタkを1増加し[ステップ406]に進む。

【００９１】

「文書検索時」

データベース検索部１０５がクラスタ文ラベルを検索するまでの動作は、第３の実施の形態と同様であるので説明は省略する。クラスタ文ラベルの検索後、データベース検索部１０５は、文書ラベル記憶部４０２から、検索された文書の文書ラベルを検索し、その結果を利用者入出力部１０６と表示部１０８を通じて利用者に表示する。検索結果の例を図２０に示す。図２０と第３の実施の形態での表示例の図１６との違いは、文書に文書ラベルである文が下線付きで表示されていることである。

【００９２】

以上のように、第４の実施の形態によれば、文書の特徴ベクトルを算出し、特徴ベクトルに基づいて文書をクラスタ分類し、クラスタの文ラベルを作成し、文書の文書ラベルを作成して記憶し、文書の検索結果をクラスタごとにまとめてクラスタ文ラベルや文書ラベルとともに表示することにより、検索結果を類似した文書の固まりとして把握し、なおかつ、固まりの内容、および、固まりに含まれる各文書の内容を理解することが容易となるので、利用者の文書検索の負担を軽減することが可能となる。

【００９３】

なお、本実施の形態では、検索結果としてクラスタID、文書IDをも表示したが、表示しなくても良い。

【００９４】

（実施の形態５）

次に本発明の第５の実施の形態について説明する。第５の実施の形態は、自由文による質問から、過去の事例から検索して適切な回答を返答する自動質問回答システムの例である。第５の実施の形態の機能構成図を図２１に示す。図２１に

おいて、図１７に示した第４の実施の形態と同じ構成部分には同一番号を付して説明は省略する。第５の実施の形態と第４の実施の形態との相違点は、特徴ベクトル抽出部５０１が抽出した特徴ベクトルを記憶する特徴ベクトル記憶部５０１と、利用者の検索質問文と特徴ベクトル記憶部５０１が記憶している文書の特徴ベクトルとの類似度を求める類似度演算部５０２とを追加したことである。

【００９５】

文書記憶部１０１には、検索の対象となる所与の複数の文書が記憶されている。記憶されている文書の例を図２２に示す。文書は、ユニークな文書ID、文章形式の質問と質問に対する回答との事例からなる。

【００９６】

以上のように構成された第５の実施の形態の動作を、第１の実施の形態と同様に、文書登録時と質問回答時に分けて以下に説明する。

【００９７】

「文書登録時」

まず、特徴ベクトル抽出部１０２は、文書記憶部１０１に記憶されている全ての文書D_iから、各文書の質問と回答のそれぞれの特徴ベクトルV_{Qi}、V_{Ai}を抽出し、抽出された特徴ベクトルを特徴ベクトル記憶部５０１に書き出す。特徴ベクトルの抽出手順は第１の実施の形態と同様であるので説明を省略する。第１の実施の形態との違いは、文書の質問と回答の部分についてそれぞれ特徴ベクトルを算出する点と、特徴ベクトルを特徴ベクトル記憶部５０１に書き出す点である。

【００９８】

次に、クラスタ分類部１０３は、特徴ベクトル記憶部５０１から回答の特徴ベクトルV_{Ai}を読み出し、全ての文書をクラスタC_kに分類し、クラスタ記憶部１０４にクラスタを書き出す。クラスタ分類の手順は第１の実施の形態と同様であるので説明を省略する。第１の実施の形態との違いは、回答の特徴ベクトルを用いてクラスタ分類を行う点である。分類されたクラスタの例として第１の実施の形態と同じ図７のものを用いる。

【００９９】

クラスタ文ラベル作成部３０１は、各クラスタの文ラベルを作成し、クラスタ

ラベル記憶部２０２に書き出す。クラスタの文ラベルの作成手順は第３の実施の形態と同様であるので説明を省略する。作成された文ラベルの例として第３の実施の形態と同じ図１５のものを用いる。

【０１００】

文書ラベル作成部４０１は、各文書の文書ラベルを作成し、文書ラベル記憶部４０２に書き出す。文書ラベルの作成手順は第４の実施の形態と同様であるので説明を省略する。作成された文書ラベルの例として第４の実施の形態と同じ図１９のものを用いる。

【０１０１】

以上の動作により、文書登録時に、質問と回答についてそれぞれ特徴ベクトルを作成し、また、回答について、クラスタ、クラスタ文ラベル、文書ラベルを作成し、それぞれの記憶部に記憶しておく。

【０１０２】

「文書検索時」

まず、利用者入出力部１０６は、入力部１０７を通じて、自由文による利用者からの質問Qを受け付ける。

【０１０３】

特徴ベクトル抽出部１０２は、質問Qの特徴ベクトルV_Qを抽出する。特徴ベクトルの抽出手順を図２３に示すフローチャートを用いて説明する。

【０１０４】

[ステップ501]　単語の抽出

質問Qから出現する単語t_jを抽出し、t_jの文書内での出現回数f_{ij}をカウントする。単語を抽出する方法は、第１の実施の形態と同様である。

【０１０５】

[ステップ502]　idf値算出

t_jのidf値を算出する。idf値として、単語t_jが文書記憶部１０１のいずれかの文書中に存在する場合は、idf_jが既に文書登録時に算出されているので、それを用いる。単語t_jが存在しない場合は（数５）により算出する。

【０１０６】

【数５】

$$idf_j = \log(n+1)+1$$

【０１０７】

[ステップ503]　tf･idf値算出

単語t_jのtf･idf値を算出する。tf･idf値算出の方法は、第１の実施の形態と同様である。抽出された質問の特徴ベクトルの例を図２４に示す。

【０１０８】

類似度演算部５０２は、特徴ベクトル記憶部５０１から全ての文書の質問の特徴ベクトルV_{Qi}を取り出し、文書の質問の特徴ベクトルV_{Qi}と利用者からの質問の特徴ベクトルV_Qとの類似度を算出する。類似度の算出手順を図２５に示すフローチャートを用いて説明する。

【０１０９】

[ステップ511]　カウンタの初期化

文書のカウンタiにi=1を設定する。

【０１１０】

[ステップ512]　特徴ベクトル内積演算

特徴ベクトルV_{Qi}と利用者からの質問の特徴ベクトルV_Qとの類似度e_iを（数６）によりベクトルの内積で算出する。

【０１１１】

【数６】

$$e_i = V_{Qi} \cdot V_Q = \frac{\sum_j w_{ij} \cdot w_{Qj}}{|V_{Qi}| \cdot |V_Q|}$$

【０１１２】

[ステップ513]　終了判定

全文書につき[ステップ512]の処理が終了した場合、すなわちi=nの場合[ステップ515]に進む。そうでない場合[ステップ514]に進む。

【０１１３】

[ステップ514]　カウンタの増加

カウンタiを1増加し[ステップ512]に進む。

【０１１４】

[ステップ515]　文書のソート

全ての文書D_iを、[ステップ512]で求めた類似度e_iの高い順にソートする。

【０１１５】

データベース検索部１０５は、文書記憶部１０１から、類似度演算部５０２が算出した類似度e_iが最上位の文書D_iを検索し、次に、クラスタ記憶部１０４から文書D_iが含まれるクラスタC_kを検索し、最後に、再び文書記憶部１０１からクラスタC_kに含まれる文書を検索し、その結果を利用者入出力部１０６に送る。

【０１１６】

利用者入出力部１０６は、表示部１０８を通じて検索結果を利用者に表示する。検索結果の例として、第４の実施の形態と同じ図２０のものを用いる。

【０１１７】

以上のように、第５の実施の形態によれば、文書の特徴ベクトルを算出し、特徴ベクトルに基づいて文書をクラスタ分類し、クラスタの文ラベルを作成し、文書の文書ラベルを作成して記憶し、利用者からの質問が入力された場合に、特徴ベクトルに基づいて類似文書を検索し、文書の検索結果をクラスタごとにまとめてクラスタ文ラベルや文書ラベルとともに表示することにより、質問に対する回答を類似した文書の固まりとして把握し、なおかつ、固まりの内容、および、固まりに含まれる各回答の内容を理解することが容易な自動質問回答システムを提供することが可能となる。

【０１１８】

なお、本実施の形態では、特徴ベクトルの類似度演算方法としてベクトルの内積を用いたが、ベクトルの類似比を用いるなど他の方法でも良い。また、検索結果としてクラスタID、文書IDをも表示したが、表示しなくても良い。

【０１１９】

【発明の効果】

第１の実施の形態によれば、文書の特徴ベクトルを算出し、特徴ベクトルに基づいて文書をクラスタ分類して記憶し、文書の検索結果をクラスタごとにまとめて表示することにより、検索結果を類似した文書の固まりとして把握することが容易となる。

【０１２０】

第２の実施の形態によれば、文書の特徴ベクトルを算出し、特徴ベクトルに基づいて文書をクラスタ分類し、単語からなるクラスタの単語ラベルを作成して記憶し、文書の検索結果をクラスタごとにまとめて単語ラベルとともに表示することにより、検索結果を類似した文書の固まりとして把握し、なおかつ、固まりの内容を理解することが容易となる。

【０１２１】

第３の実施の形態によれば、文書の特徴ベクトルを算出し、特徴ベクトルに基づいて文書をクラスタ分類し、文からなるクラスタの文ラベルを作成して記憶し、文書の検索結果をクラスタごとにまとめて文ラベルとともに表示することにより、検索結果を類似した文書の固まりとして把握し、なおかつ、固まりの内容を自然文として理解することが容易となる。

【０１２２】

第４の実施の形態によれば、文書の特徴ベクトルを算出し、特徴ベクトルに基づいて文書をクラスタ分類し、クラスタの文ラベルを作成し、文書の文書ラベルを作成して記憶し、文書の検索結果をクラスタごとにまとめてクラスタ文ラベルや文書ラベルとともに表示することにより、検索結果を類似した文書の固まりとして把握し、なおかつ、固まりの内容、および、固まりに含まれる各文書の内容を理解することが容易となる。

【０１２３】

第５の実施の形態によれば、文書の特徴ベクトルを算出し、特徴ベクトルに基づいて文書をクラスタ分類し、クラスタの文ラベルを作成し、文書の文書ラベルを作成して記憶し、利用者からの質問が入力された場合に、特徴ベクトルに基づいて類似文書を検索し、文書の検索結果をクラスタごとにまとめてクラスタ文ラベルや文書ラベルとともに表示することにより、質問に対する回答を類似した文

書の固まりとして把握し、なおかつ、固まりの内容、および、固まりに含まれる各回答の内容を理解することが容易な自動質問回答システムを提供することが可能となる。

【図面の簡単な説明】

【図１】

本発明第１の実施の形態の機能構成図

【図２】

本発明第１の実施の形態のハードウェア構成を示す図

【図３】

本発明第１の実施の形態の文書の例を示す図

【図４】

本発明第１の実施の形態の文書の特徴ベクトルの例を示す図

【図５】

本発明第１の実施の形態の特徴ベクトルの抽出手順を示すフローチャート

【図６】

本発明第１の実施の形態のクラスタ分類手順を示すフローチャート

【図７】

本発明第１の実施の形態のクラスタの例を示す図

【図８】

本発明第１の実施の形態の検索結果の例を示す図

【図９】

本発明第２の実施の形態の機能構成図

【図１０】

本発明第２の実施の形態のクラスタ単語ラベルの作成手順を示すフローチャート

【図１１】

本発明第２の実施の形態のクラスタ単語ラベルの例を示す図

【図１２】

本発明第２の実施の形態の検索結果の例を示す図

【図１３】

本発明第３の実施の形態の機能構成図

【図１４】

本発明第３の実施の形態のクラスタ文ラベルの作成手順を示すフローチャート

【図１５】

本発明第３の実施の形態のクラスタ文ラベルの例を示す図

【図１６】

本発明第３の実施の形態の検索結果の例を示す図

【図１７】

本発明第４の実施の形態の機能構成図

【図１８】

本発明第４の実施の形態の文書ラベルの作成手順を示すフローチャート

【図１９】

本発明第４の実施の形態の文書ラベルの例を示す図

【図２０】

本発明第４の実施の形態の検索結果の例を示す図

【図２１】

本発明第５の実施の形態の機能構成図

【図２２】

本発明第５の実施の形態の文書の例を示す図

【図２３】

本発明第５の実施の形態の利用者の質問の特徴ベクトルの抽出手順を示すフローチャート

【図２４】

本発明第５の実施の形態の利用者の質問の特徴ベクトルの例を示す図

【図２５】

本発明第５の実施の形態の類似度演算の手順を示すフローチャート

【符号の説明】

１０１　文書記憶部

１０２　特徴ベクトル抽出部

１０３　クラスタ分類部

１０４　クラスタ記憶部

１０５　データベース検索部

１０６　利用者入出力部

１０７　入力部

１０８　表示部

１１１　ＣＰＵ

１１２　主記憶装置

１１３　補助記憶装置

２０１　クラスタ単語ラベル作成部

２０２　クラスタラベル記憶部

３０１　クラスタ文ラベル作成部

４０１　文書ラベル作成部

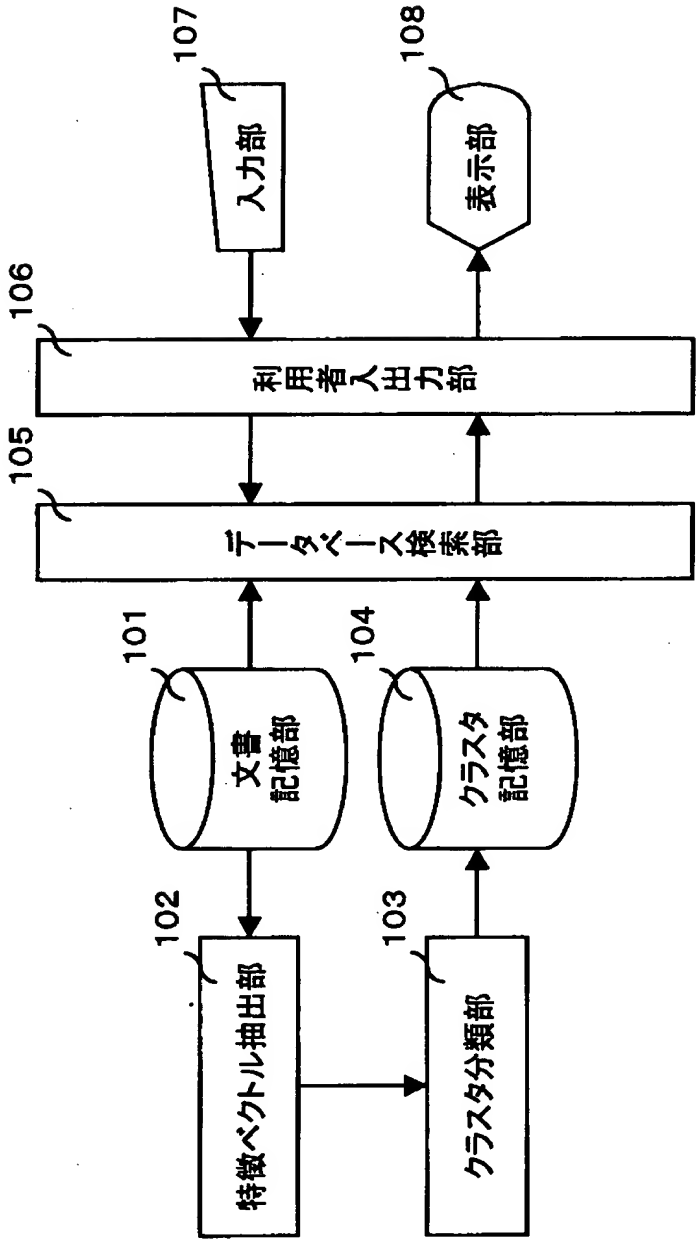
４０２　文書ラベル記憶部

５０１　特徴ベクトル記憶部

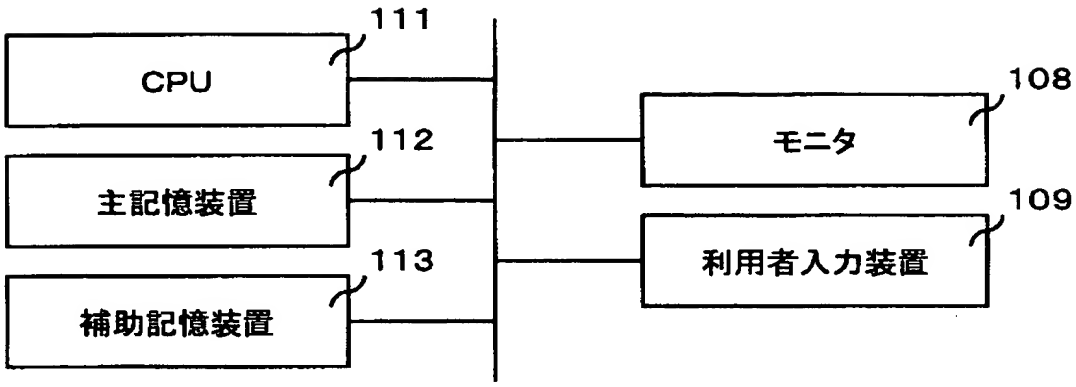
５０２　類似度演算部

【書類名】　　図面

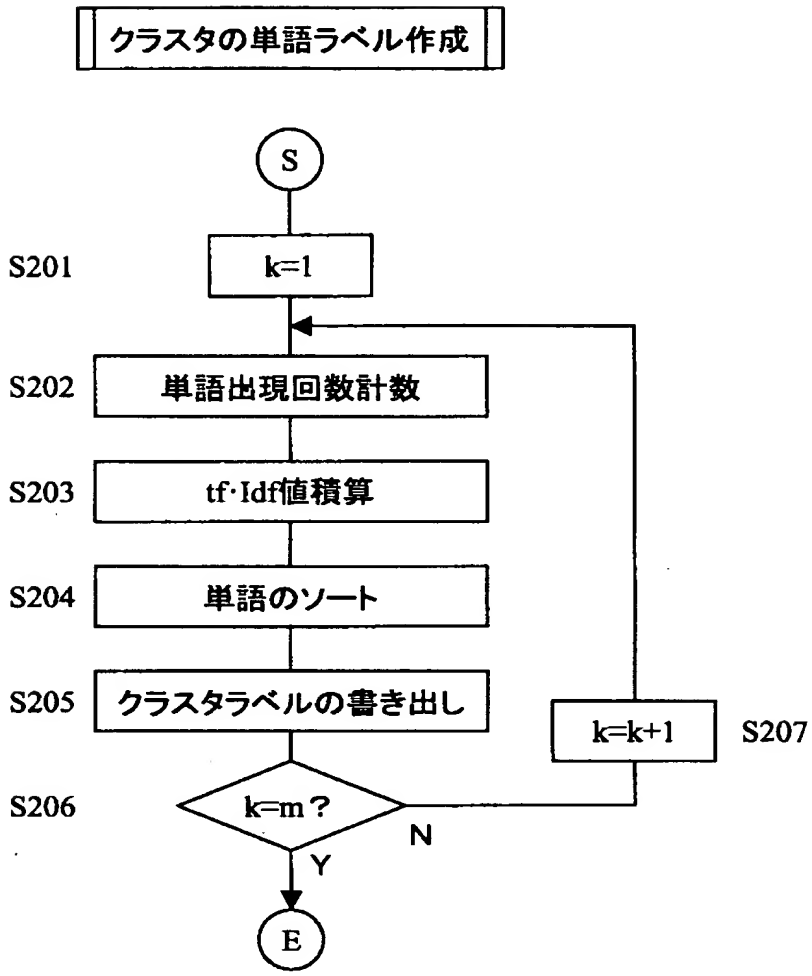
【図１】


107
入力部
108
表示部
106
利用者入出力部
105
データベース操作部
101
文書記憶部
104
クラスタ記憶部
102
特徴ベクトル抽出部
103
クラスタ分類部

【図２】



【図１０】



【図１１】

クラスタID	単語ラベル
1	お菓子, 間食, チーズ
2	体調, エクササイズ, 効果
3	ストレス, 前向き, 状況
4	生理, 食欲, ダイエット
:	:

【図１２】

○○件の検索結果がありました。

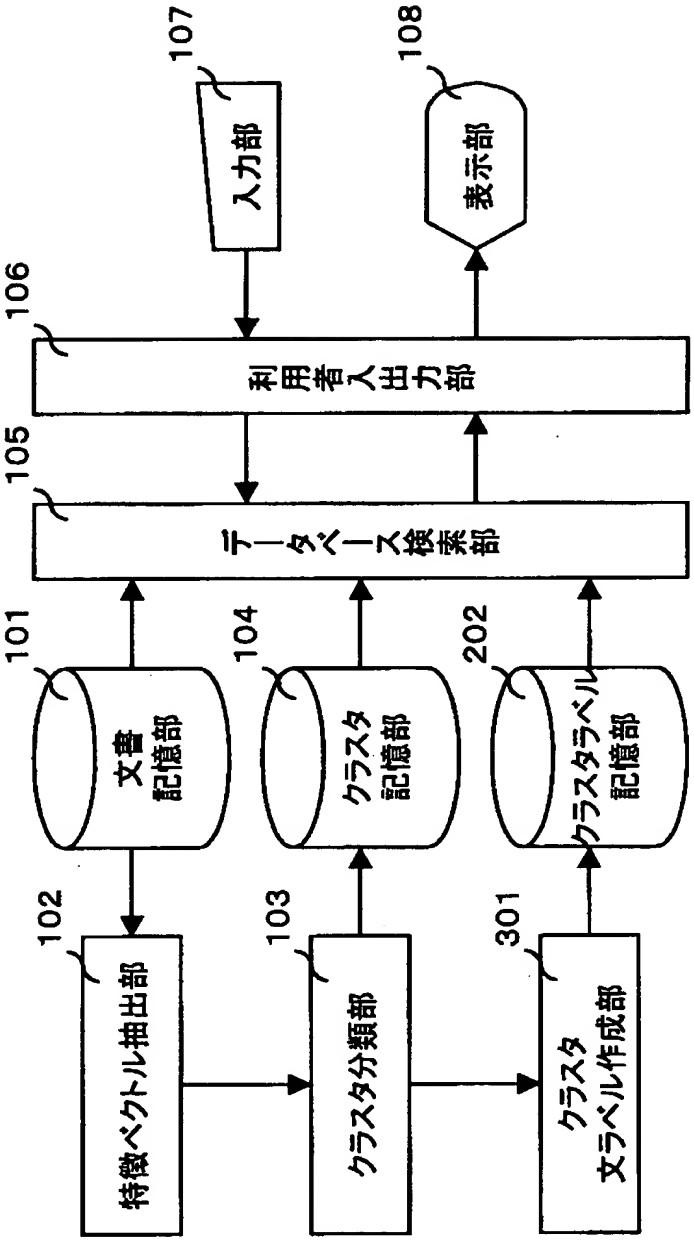
クラスタID	単語ラベル	文書ID	文書
1	お菓子 間食 チーズ	1	お菓子が食べたくなったときは、 ①かみごたえがあり、後を引かないもので、量を決めて食べる。・・・
		190	間食には、牛乳、乳製品(チーズ・ヨーグルトなど)、・・・
		432	間食は200kcal以内で自由に選んでもOKです。・・・
		644	お菓子が食べたくなったときは、 ・かみごたえがあり、後を引かないもので、・・・

前のクラスタ　　次のクラスタ

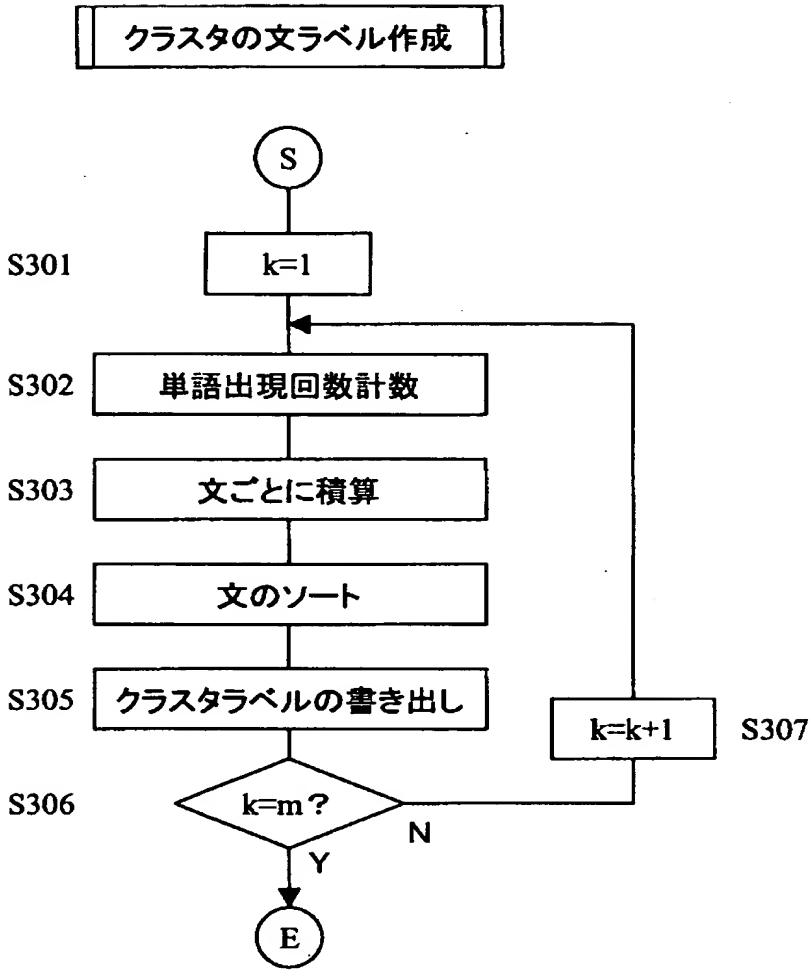
特２０００－１４５１６８

出証特２００１－３０４１５４５

【図１３】


107
入力部
108
表示部
106
利用者入出力部
105
データベース操作部
101
文書記憶部
104
クラスタ記憶部
202
クラスタラベル記憶部
102
特徴ベクトル抽出部
103
クラスタ分類部
301
クラスタ文ラベル作成部

【図１４】



【図１５】

クラスタID	文ラベル
1	水分の多い物（ゼリー、プリン、ヨーグルト）を・・・
2	生理中、体調が悪い時には無理にエクササイズを・・・
3	何がストレスになっているのかを確かめ、・・・
4	そこで生理前でも自分のダイエットペースを崩さないように・・・
：	：

【図１６】

○○件の検索結果がありました。

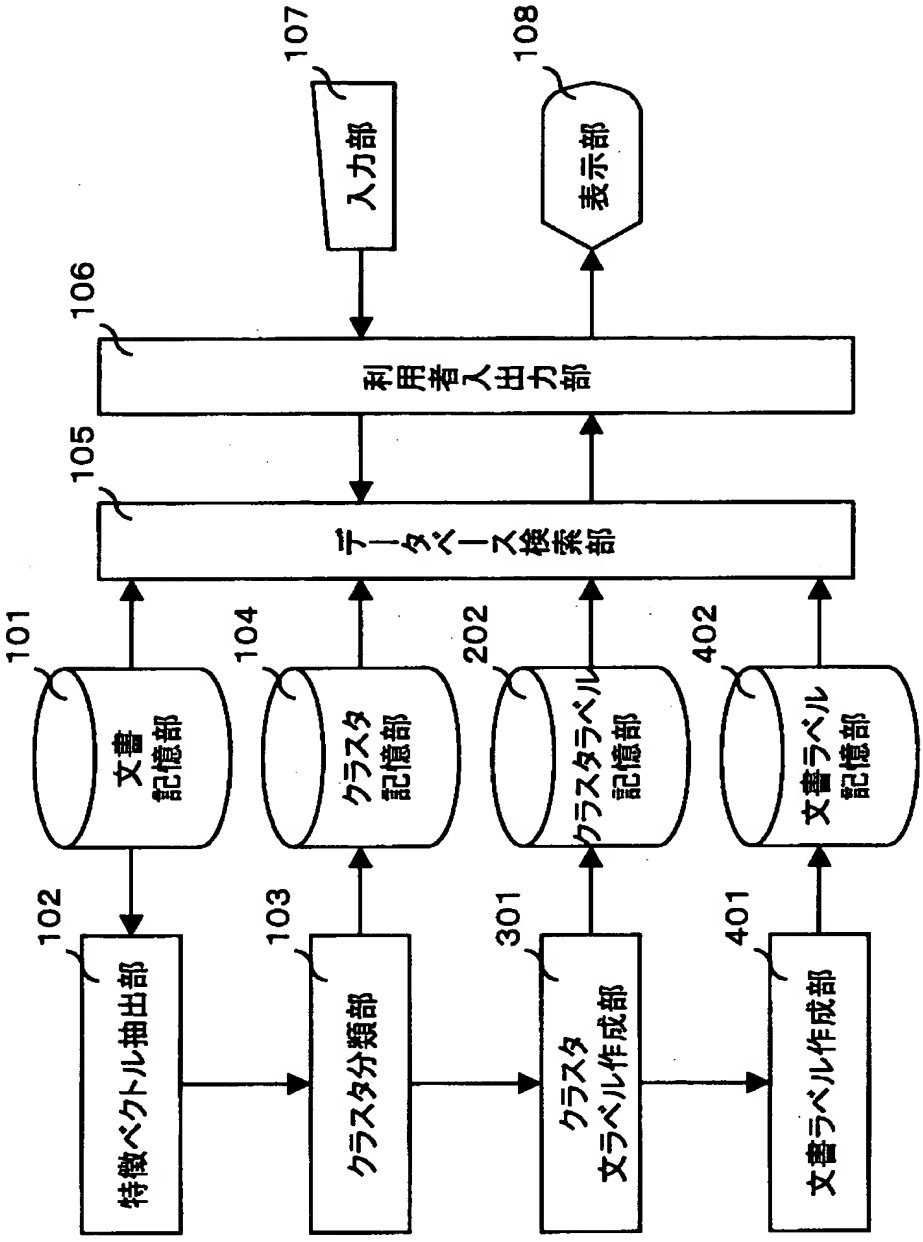
クラスタID	文ラベル	文書ID	文書
1	水分の多い物(ゼリー、プリン、ヨーグルト)を・・・	1	お菓子が食べたくなったときは、 ① かみごたえがあり、後を引かないもので、量を決めて食べる。・・・
		190	間食には、牛乳、乳製品(チーズ・ヨーグルトなど)、・・・
		432	間食は200kcal以内で自由に選んでもOKです。・・・
		644	お菓子が食べたくなったときは、 ・かみごたえがあり、後を引かないもので、・・・

前のクラスタ　　次のクラスタ

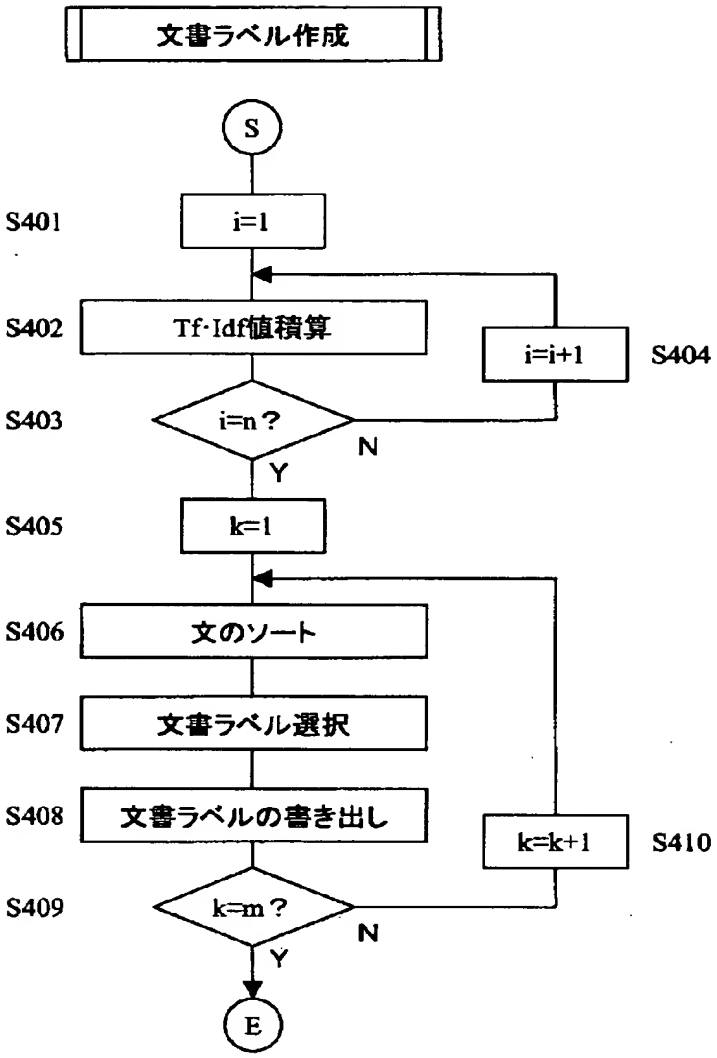
特２０００－１４５１６８

出証特２００１－３０４１５４５

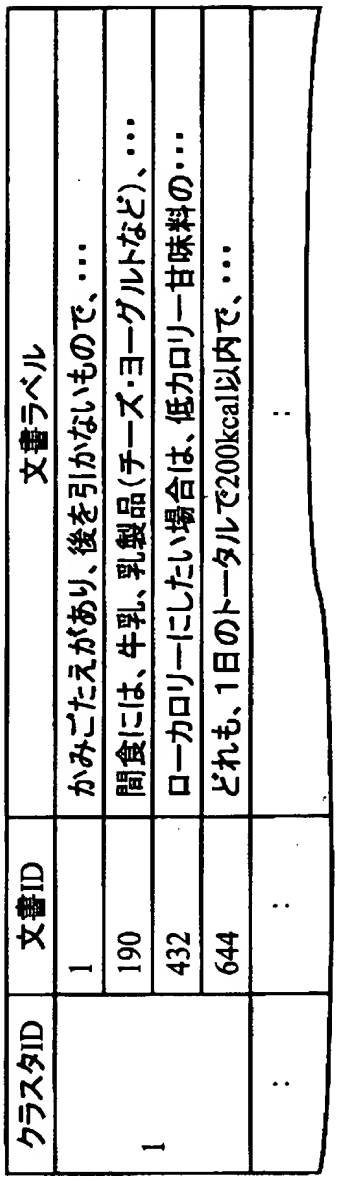
【図１７】


107
入力部
108
表示部
106
利用者入出力部
105
データベース検索部
101
文書記憶部
104
クラスタ記憶部
202
クラスタラベル記憶部
402
文書ラベル記憶部
102
特徴ベクトル抽出部
103
クラスタ分類部
301
クラスタ文ラベル作成部
401
文書ラベル作成部

【図１８】


文書ラベル作成
S
S401
i=1
S402
Tf･Idf値積算
i=i+1
S404
S403
i=n？
N
Y
S405
k=1
S406
文のソート
S407
文書ラベル選択
S408
文書ラベルの書き出し
k=k+1
S410
S409
k=m？
N
Y
E

【図１９】



クラスタID	文書ID	文書ラベル
1	1	かみごたえがあり、後を引かないもので、・・・
	190	間食には、牛乳、乳製品（チーズ・ヨーグルトなど）、・・・
	432	ローカロリーにしたい場合は、低カロリー甘味料の・・・
	644	どれも、１日のトータルで200kcal以内で、・・・
：	：	：

【図２０】

○○件の検索結果がありました。

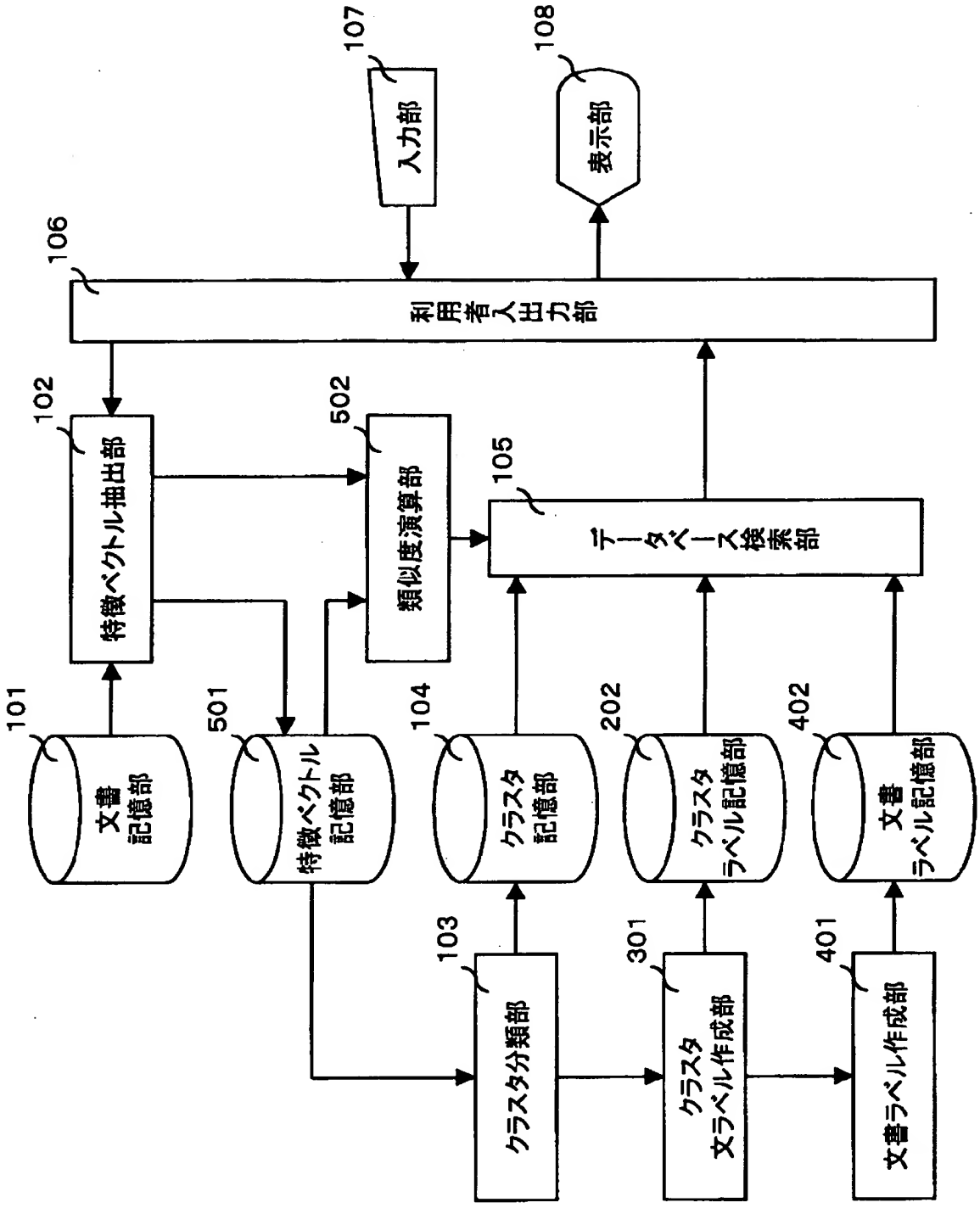
クラスタID	文ラベル	文書ID	文書
1	水分の多い物(ゼリー、プリン、ヨーグルト)を・・・	1	お菓子が食べたくなったときは、 ①かみごたえがあり、後を引かないもので、量を決めて食べる。・・・
		190	間食には、牛乳、乳製品(チーズ・ヨーグルトなど)、・・・
		432	間食は200kcal以内で自由に選んでもOKです。・・・ ローカロリーにしたい場合は、低カロリー甘味料の・・・
		644	お菓子が食べたくなったときは、 ・かみごたえがあり、後を引かないもので、・・・ どれも、1日のトータルで200kcal以内で、・・・

前のクラスタ　　次のクラスタ

出証特２００１－３０４１５４５

特２０００－１４５１６８

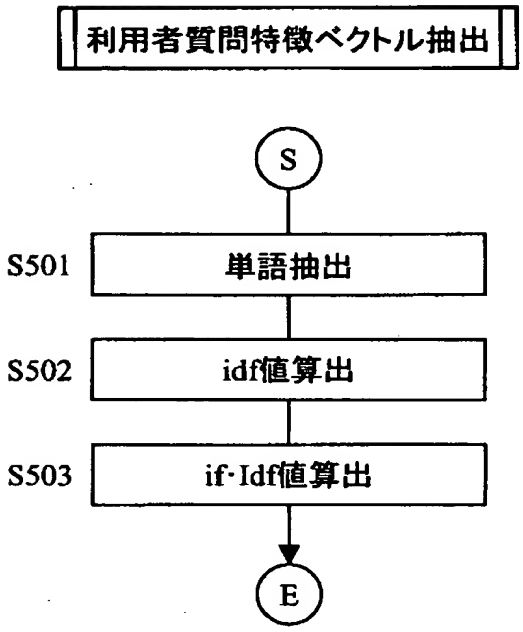
【図２１】


107
入力部
108
表示部
106
利用者入出力部
102
特徴ベクトル抽出部
502
類似度演算部
105
データベース検索部
101
文書記憶部
501
特徴ベクトル記憶部
104
クラスタ記憶部
202
クラスタラベル記憶部
402
文書ラベル記憶部
103
クラスタ分類部
301
クラスタ文ラベル作成部
401
文書ラベル作成部

【図２２】

文書ID	質問	回答
1	どうしても、お菓子が食べたくなったら、どんなものを食べたらいいですか。…	お菓子が食べたくなったときは、 ①かみごたえがあり、後を引かないもので、量を決めて食べる。…
2	朝一番の食事の前の運動が一番効果があると聞いたのですが、夕方の有酸素運動はどうなんですか？…	運動はいつでも、どこでも一人で、が原則です。 人によって生活、体調はみな違います。…
3	今月の生理が１週間以上も遅れてしまってます。…	まったくあなたの考えている通りです。何か悩みがあったり、…
： n個	：	：

【図２３】

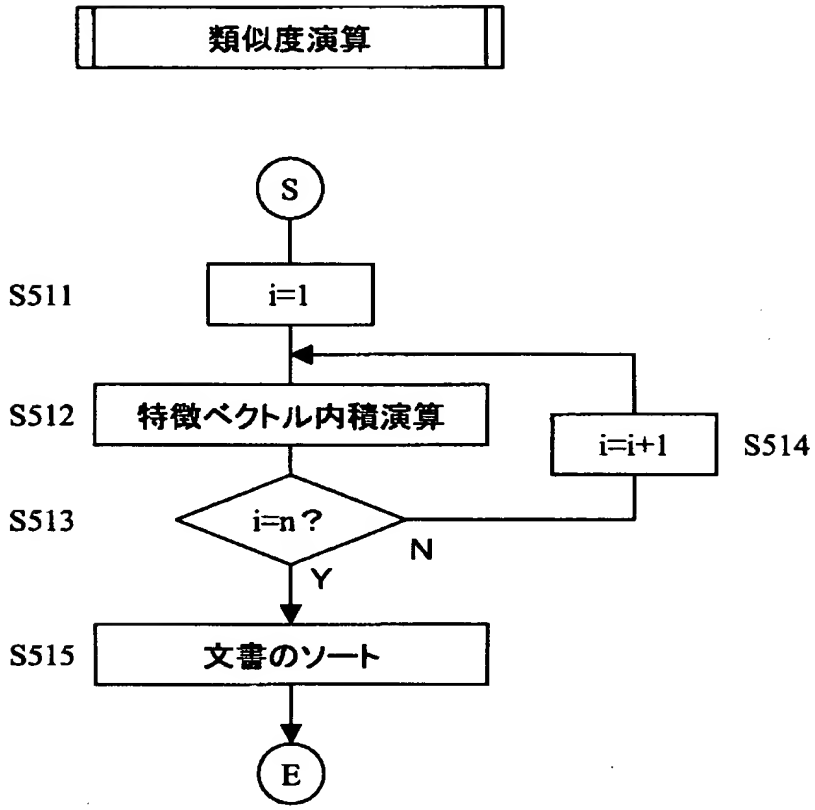


【図２４】

特徴ベクトル V_Q

単語 t_i	重み w_{Qi}
お菓子	0.601
間食	0.452
私	0.400
食べる	0.847
朝食	0.556
方法	0.456
:	:

【図２５】


類似度演算
S
S511
i=1
S512
特徴ベクトル内積演算
i=i+1
S514
S513
i=n？
N
Y
S515
文書のソート
E

【書類名】　要約書

【要約】

【課題】　従来、情報検索装置や自動質問回答装置では、質問文に類似した文書として、同じような内容の文書が大量に検索されることが多く、利用者は同じような内容の検索結果の中から欲しい情報を見つけ出す必要があったり、あるいは、同じような内容の検索結果が大量に出てきても、自分の欲しい情報がなかったりするという課題があった。

【解決手段】　情報単位の特徴量を抽出し、前記特徴量に基づき前記情報単位のクラスタ分類を行い、前記情報単位から検索条件に合致する情報単位を検索し、検索結果を前記クラスタにまとめて出力する。

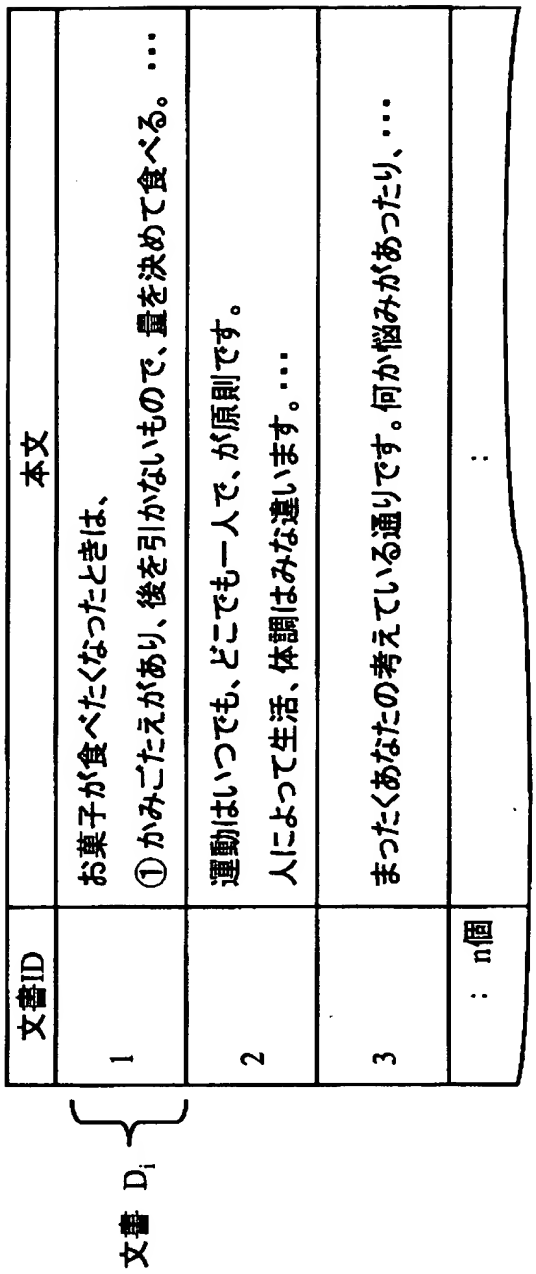
【選択図】　図１

出　願　人　履　歴　情　報

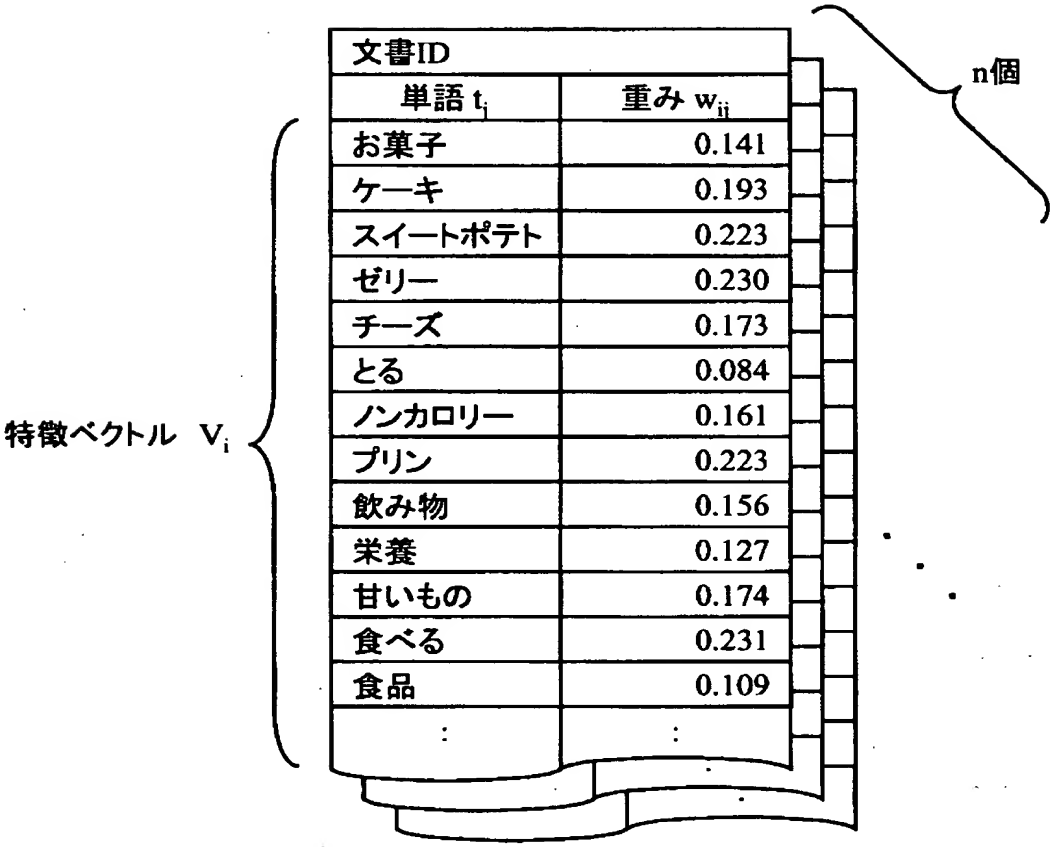
識別番号　　〔０００００５８２１〕

１．変更年月日　　１９９０年　８月２８日
［変更理由］　新規登録
住　所　　大阪府門真市大字門真１００６番地
氏　名　　松下電器産業株式会社

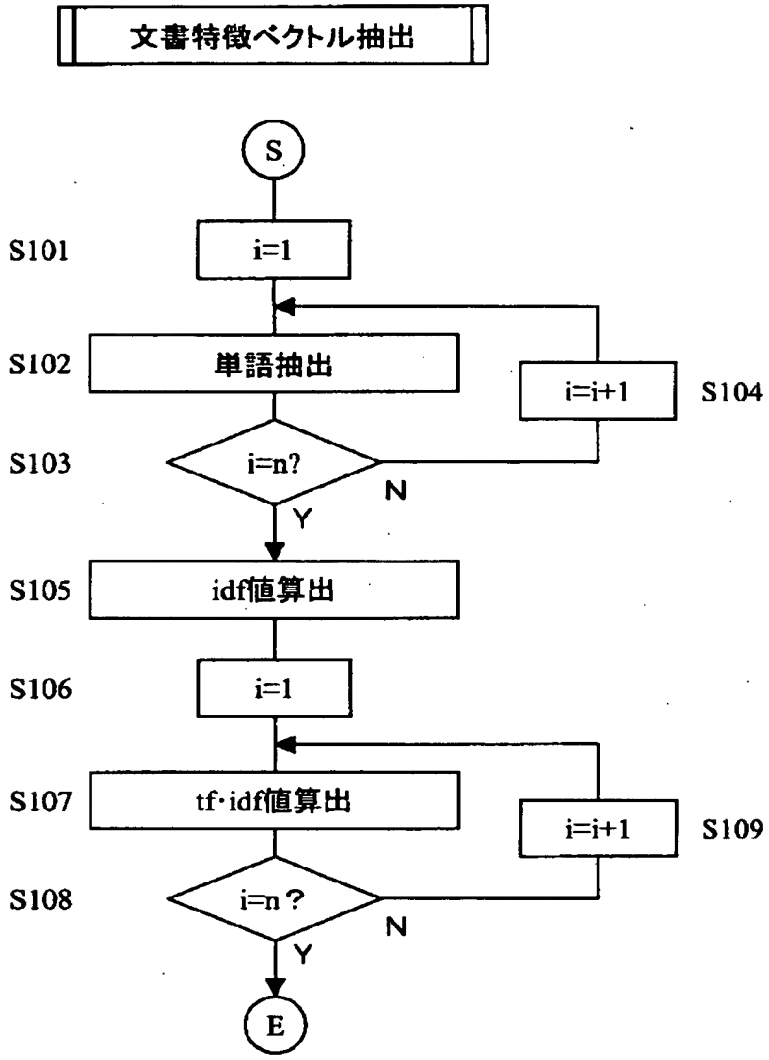
【図３】


文書 D_i
文書ID
本文
1
お菓子が食べたくなったときは、
①かみごたえがあり、後を引かないもので、量を決めて食べる。・・・
2
運動はいつでも、どこでも一人で、が原則です。
人によって生活、体調はみな違います。・・・
3
まったくあなたの考えている通りです。何か悩みがあったり、・・・
：n個
：

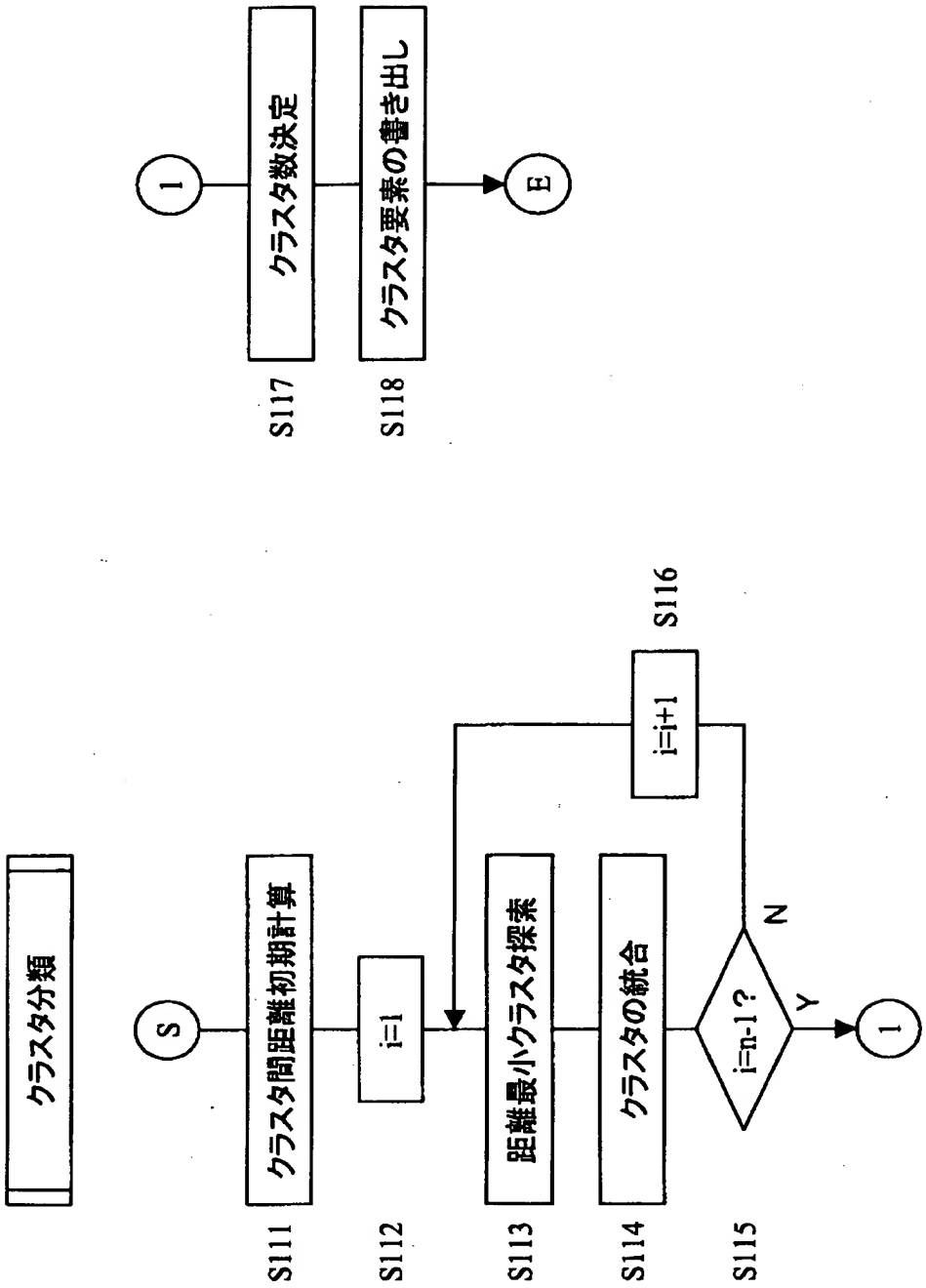
【図４】


文書ID
単語 t_j
重み w_{ij}
お菓子 0.141
ケーキ 0.193
スイートポテト 0.223
ゼリー 0.230
チーズ 0.173
とる 0.084
ノンカロリー 0.161
プリン 0.223
飲み物 0.156
栄養 0.127
甘いもの 0.174
食べる 0.231
食品 0.109
n個
特徴ベクトル V_i

【図５】


文書特徴ベクトル抽出
S
S101
i=1
S102
単語抽出
i=i+1
S104
S103
i=n?
N
Y
S105
idf値算出
S106
i=1
S107
tf･idf値算出
i=i+1
S109
S108
i=n？
N
Y
E

【図６】


クラスタ分類
S
クラスタ間距離初期計算
S111
i=1
S112
距離最小クラスタ探索
S113
クラスタの統合
S114
i=n-1？
S115
N
Y
i=i+1
S116
1
クラスタ数決定
S117
クラスタ要素の書き出し
S118
E

【図７】

クラスタID	文書ID
1	1, 190, 432, 644
2	2, 412, 3, 158
3	3, 158
4	4, 109, 182, 615
:	:

【図８】

○○件の検索結果がありました。

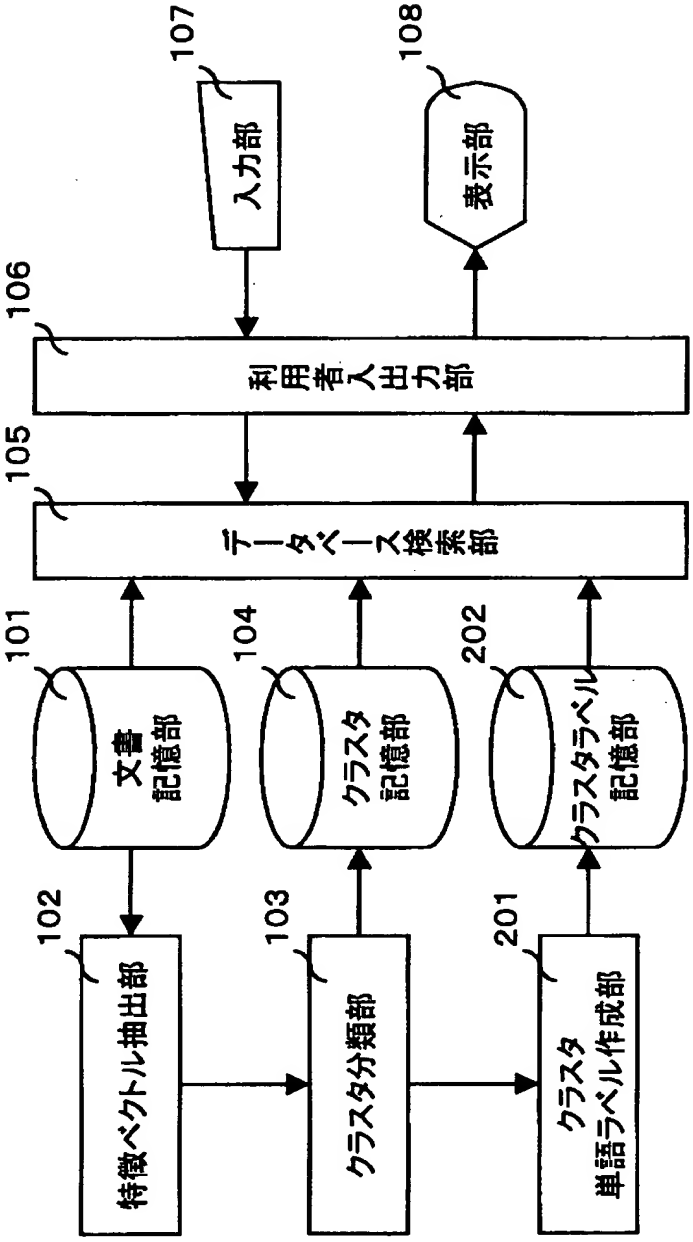
クラスタID	文書ID	文書
1	1	お菓子が食べたくなったときは、 ①かみごたえがあり、後を引かないもので、量を決めて食べる。・・・
	190	間食には、牛乳、乳製品(チーズ・ヨーグルトなど)、・・・
	432	間食は200kcal以内で自由に選んでもOKです。・・・
	644	お菓子が食べたくなったときは、 ・かみごたえがあり、後を引かないもので、・・・

前のクラスタ　　次のクラスタ

特２０００－１４５１６８

出証特２００１－３０４１５４５

【図９】


107
入力部
108
表示部
106
利用者入出力部
105
データベース操作部
101
文書記憶部
104
クラスタ記憶部
202
クラスタラベル記憶部
102
特徴ベクトル抽出部
103
クラスタ分類部
201
クラスタ単語ラベル作成部